Actively Learning $\mathcal{EL}$ Terminologies from Large Language Models

Matteo Magnini[*] Riccardo Squarcialupi[*]
Martin T. Sterri[†] Ana Ozaki[†,‡]

[*]Alma Mater Studiorum – University of Bologna
matteo.magnini@unibo.it, riccard.squarcialupi@studio.unibo.it

[†]University of Bergen
martin.sterri@student.uib.no, ana.ozaki@uib.no

[‡]University of Oslo
anaoz@ifi.uio.no

The European Conference on Artificial Intelligence (ECAI 2025)
27 October, 2025, Bologna

Context I

The active learning framework:

- a learner attempts to learn some kind of knowledge;
- by posing questions to a teacher;

- questions made by the learner are
    - membership queries → ask whether concept inclusions are true or false;
    - equivalence queries → ask whether the idea od the learner about the knowledge of the teacher is correct or not.

Context II

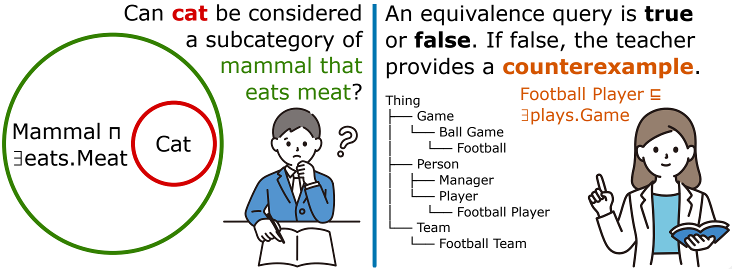


Figure: Example of membership and equivalence queries

We want to use Large Language Models (LLMs) as teachers in the Angluin's exact learning framework [Angluin, 1987].

Motivation

Motivations for our work:

- to the best of our knowledge, the only implementation of the Angluin's exact learning framework uses a synthetic teacher [Duarte et al., 2018];
- ontology construction is a costly and time-consuming task that requires domain experts;
- arguably, a boring and repetitive task for humans;
- with LLMs as teachers, we can automate the process of ontology construction;
- with Angluin's framework, we build ontologies in a systematic way.

Algorithm I

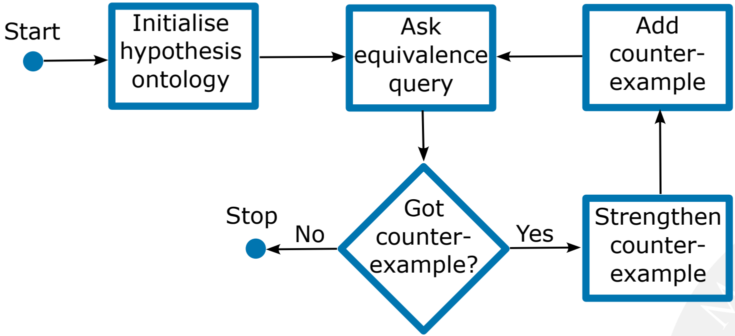


Figure: Overview of the exact learning algorithm.

Algorithm II

Equivalence query are symulated via random sampling. The algorithm checks if the classification of the examples match with the information in the hypothesis:

- true inclusions must be logical consequences;
- false ones must not.

If the hypothesis fits the classification of the concept inclusions, learning stops. Otherwise, the inclusion not fitting the hypothesis is used as a counterexample.

Algorithm III

The sampling-based simulation can yield PAC [Valiant, 1984] guarantees when the sample size

$$|S| \geq \frac{\ln\left(|H|/\gamma\right)}{\epsilon}$$

is computed from the hypothesis space H ($\mathcal{EL}$ terminologies of bounded structure) and parameters ϵ (error) and γ (confidence).

Learner's operations I

When the teacher replies with a counterexample, the learner before adding it to the hypothesis processes it. The learner performs operations, that use membership queries, in order to maximise how informative the concept inclusions are and also to minimise their size.

- Decompose Left
- Decompose Right
- Merging
- Branching
- Saturation
- Desaturation

Learner's operations II

Decompose Right

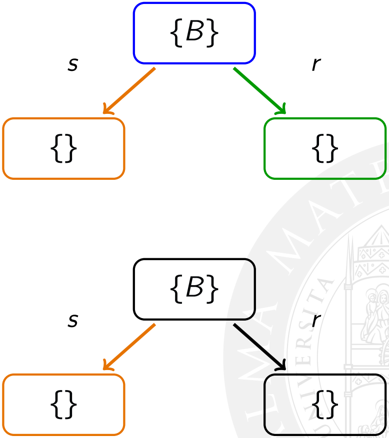
$$T = \{A \sqsubseteq \exists r.\top,\ B \sqsubseteq \exists r.\top,\ A \sqsubseteq B\}$$

$$H = \{A \sqsubseteq B\}$$

$$\mathrm{C} = A \sqsubseteq B \sqcap \exists s.\top \sqcap \exists r.\top$$

$$\Downarrow$$

$$\mathrm{C} = B \sqsubseteq \exists s.\top$$



Ontologies I

Ontology	N_C	N_R	Log. Ax.	PAC Sample	Poss. Ax.
Animals	17	4	12	542	6,936
Cell	22	0	24	1,119	10,164
Football	10	3	9	341	1,500
Generations	20	4	18	847	10,800
University	7	3	4	139	588

Table: Ontology statistics and PAC sample sizes with $\epsilon = 0.2$ and $\gamma = 0.1$. N_C and N_R are the number of concept and role names occurring in the ontologies.

Ontologies II

Ontology	N_C	N_R	Log. Ax.	PAC Sample	Pos. Ax.
Ab. Elb. J. C.	27	14	43	2,286	39,366
BNF Sec.	36	24	80	4,646	107,568
Chlorhexidine	23	14	38	1,946	26,450
Cone of Tissue	42	42	100	6,163	220,500
Kalli Krein	18	10	27	1,279	11,988
Neon	16	10	25	1,149	8,960
Pin	43	40	99	6,113	225,578
Pros. Drug	29	14	47	2,540	47,096
Zopiclone	32	36	77	4,465	105,472
Zuccini	33	22	58	3,295	82,764

Table: Ontology statistics and PAC sample sizes with $\epsilon = 0.2$ and $\gamma = 0.1$ for medical ontologies (sub modules of the Galen ontology [Alan L. Rector, 1996]).

LLMs and how to query them I

$\text{Cat} \sqsubseteq \text{Mammal} \sqcap \exists \text{eats.Meat}$

- Manchester OWL Syntax
  - Cat SubClassOf Mammal and eats some Meat?
- Natural Language
  - Can Cat be considered a subcategory of "Mammal that is also something that eats some Meat"?

LLMs used as teachers:

- Llama2 (13B)
- Llama3 (8B)
- Mistral (7B)
- Mixtral (47B)

LLMs and how to query them II

Two different system prompts used to query the LLMs:

- Concise:
  - Answer with only True or False.
- Detailed:
  - You need to classify the following statements as True or False. The statement will be provided in either Manchester OWL syntax or natural language. Strictly follow these guidelines:
    1. answer with only True or False;
    2. entities with has part relation are not in a subclass relation;
    3. take a deep breath before answering;
    4. if you are unsure about the classification, answer with False.

Evalualtion I

The metrics are computed considering all possible axioms of the form:

- $A \sqsubseteq B$
- $A \sqcap B \sqsubseteq C$
- $B \sqsubseteq \exists r.A$
- $\exists r.A \sqsubseteq B$

These axioms are formulated with a finite signature. Tautologies, such as:

- $A \sqsubseteq A$
- $A \sqcap B \sqsubseteq B$
- $A \sqcap B \sqsubseteq A$

are removed to avoid artificially inflating true positives.

Evalualtion II

Axioms are classified as:

- TP Entailed by both the original and learnt ontology;
- TN Not entailed by either ontology;
- FN Entailed by the original ontology but not the learnt one;
- FP Entailed by the learnt ontology but not the original one.

Results I

Ontology	Accuracy	Recall	Precision	F1-Score
Animals	0.737	0.858	0.381	0.428
Cell	0.391	0.733	0.206	0.284
Football	0.553	0.89	0.422	0.477
Generations	0.691	0.658	0.564	0.476
University	0.622	0.629	0.313	0.302

Table: Results of ExactLearner+LLM grouped by ontologies.

Model	Accuracy	Recall	Precision	F1-Score
Llama2 (13b)	0.521	0.71	0.294	0.314
Llama3 (8b)	0.43	0.947	0.218	0.333
Mistral (7b)	0.741	0.747	0.45	0.49
Mixtral (47b)	0.705	0.611	0.547	0.436

Table: Results of ExactLearner+LLM grouped by models.

Results II

Prompt Type	Accuracy	Recall	Precision	F1-Score
M. OWL Syntax	0.34	0.93	0.165	0.262
Natural Language	0.751	0.811	0.414	0.511
A. M. OWL Syntax	0.537	0.767	0.326	0.347
A. Natural Language	0.767	0.506	0.603	0.454

Table: Results of ExactLearner+LLM grouped by prompts.

Results III

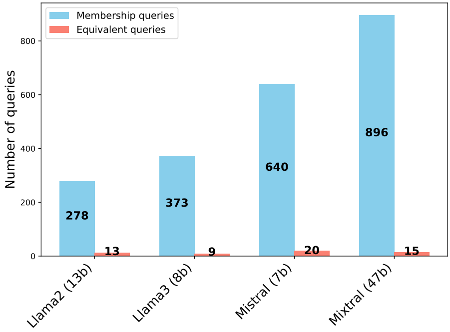


Figure: Average number of membership and (simulated) equivalence queries grouped by LLM.

Results IV

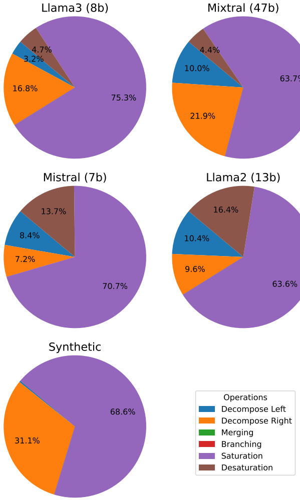


Figure: Aggregated results of the operations performed by the learner during the PAC learning of all the ontologies grouped by teacher type (LLMs and synthetic).

Actively Learning $\mathcal{EL}$ Terminologies from Large Language Models

Matteo Magnini[*] Riccardo Squarcialupi[*]
Martin T. Sterri[†] Ana Ozaki[†,‡]

[*]Alma Mater Studiorum – University of Bologna
matteo.magnini@unibo.it, riccard.squarcialupi@studio.unibo.it

[†]University of Bergen
martin.sterri@student.uib.no, ana.ozaki@uib.no

[‡]University of Oslo
anaoz@ifi.uio.no

The European Conference on Artificial Intelligence (ECAI 2025)
27 October, 2025, Bologna

References I

[Alan L. Rector, 1996] Alan L. Rector, J.E. Rogers, P. P. (1996).
The galen high level ontology.
In *Medical Informatics Europe*, page 174–178. IOS Press
DOI:10.3233/978-1-60750-878-6-174.

[Angluin, 1987] Angluin, D. (1987).
Queries and concept learning.
Mach. Learn., 2(4):319–342
DOI:10.1007/BF00116828.

[Duarte et al., 2018] Duarte, M. R. C., Konev, B., and Ozaki, A. (2018).
Exactlearner: A tool for exact learning of EL ontologies.
In Thielscher, M., Toni, F., and Wolter, F., editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018*, pages 409–414. AAAI Press
https://aaai.org/ocs/index.php/KR/KR18/paper/view/18006.

References II

[Valiant, 1984] Valiant, L. G. (1984).
A theory of the learnable.
Commun. ACM, 27(11):1134–1142
DOI:10.1145/1968.1972.